ON COUNTING MEANINGFUL UNITS IN TEXTS

Maurice Gross

ON COUNTING MEANINGFUL UNITS IN TEXTS

Maurice Gross

Université Paris 7
Laboratoire d'Automatique Documentaire et Linguistique
2, place Jussieu, 75005 Paris, France

L'analyse syntaxique automatique, première étape d'une procédure d'interprétation fine des textes par ordinateur, a recours à des outils comme les grammaires et les dictionnaires. Ces outils, tels qu'ils sont actuellement disponibles, ne sont pas suffisants. Ils doivent en effet prendre une forme électronique qui impose des révisions majeures de leur forme et contenu. Nous présentons une méthodologie linguistique qui a permis de construire des outils électroniques à large couverture des langues. Ces nouveaux outils mettent en évidence des unités linguistiques signifiantes, ce qui conduit à une modification substantielle de l'analyse du contenu des textes.

KEY WORDS Electronic dictionaries, Electronic Grammars, Parsing, Compound Words, Units of Meaning, Corpus Analysis.

Today, all texts are produced by computers: large technical documentation, books, journals, newspapers, commercial and personal letters, etc. As a consequence, texts are stored on a magnetic support and can be archived and retrieved by computer from this medium. This situation is entirely new: less than ten years ago, in order to process a text by computer, it had to be typed in manually from its paper version. The amount of material obtainable in this way represented a considerable limitation, both qualitative and quantitative, on the requirements of automatic text processing. Large amounts of texts are only now being made available to the public, mainly newspapers on CD-ROMs. Studies that could not even begin because of the lack of primary material can now be undertaken.

Automatic processing tools are applied to this raw material, namely: texts as made up of sequences of words[1]. We discuss first this point of departure.

1. The notion of word

A formal notion of word is relatively easy to define. Thus, at first sight, the words of a text are sequences of characters separated by a space. More precisely, a full list of separators of words can be drawn up for the printed texts of a given language, such a list will include more complex separators than the space: combinations of spaces and of punctuation marks, special characters such as the hyphen, the apostrophe, commas in numerals, etc[2].

It is not always perceived that the separation of written words as taught today for

[1] Basically, texts are sequences of ASCII characters structured as strings of words by means of a set of separators, themselves ASCII characters. We will not deal here with this alphabetic or possibly phonetic level.

[2] For texts on a computer support, the list may have to be extended so as to include formatting characters specific to a given word processor.

European languages is the latest stage of a complex process of linguistic analysis that has been at work over several centuries. Because of the historical nature of the process, segmentation of texts into words is by no means exempt of contradictions and other irrational features. We will illustrate this situation through a few examples.

Consider the clitic pronouns in French and in Italian. Clitics are words, since they are strings of letters separated by a space or an apostrophe, as in:

Il me l'a donné *Me lo ha dato*

However, in the imperative, the situation is different, we have:

Donne-le-moi *Damelo*

In French, the separators are now hyphens and no space is involved; in Italian, the same clitics are no longer formal words, since they are merged with their verb. In French, we can maintain the status of words for clitics, just by putting the hyphen in the list of separators[3]. This cannot be done in Italian, for recovering the clitics from the single word *damelo* requires a complex morpho-syntactic analysis[4]. In the same way, reflexive clitics are words in French as in *se laver*, but not in the corresponding Italian form: *lavarse*.

Languages such as Finnish and Hungarian have affixes similar to the above mentioned clitics. Their spelling tradition attaches these particles to the root words. Notice that any other tradition could have been initiated by the early grammarians of these languages. For example, particles could have been separated by hyphens or by spaces, thus clearly isolating the roots, hence making much easier a dictionary look up for persons having no a priori knowledge of the language.

Similar remarks apply to German compounds:

- the separable particles of verbs could well have been isolated as words:

aus-wahlen and *aus-gewahlt* instead of *auswahlen* and *ausgewahlt*

- compound words could have been divided and boundaries between components marked by hyphens: *Herz-klopfen* instead of *Herzklopfen.* In that way, the component strings would have kept their status of simple word that they have in other contexts.

Actually, French presents a situation similar to that of German for certain compounds whose spelling is not standardized, such as the two attested spellings *deltaplane* and *delta-plane* for the same 'word', that is, the same semantic unit. This is also true for many compounds which include forms suffixed in *-o*: *microéconomie* and *micro-économie*. Major dictionary publishers and the French Academy constitute together an institutional body which has succeeded so far in maintaining standards of spelling for simple words. But this institution has neither power nor influence on the domain of compound words[5], hence all possible forms are found in texts.

Tenses in many languages provide other difficulties of counting procedures. The preterit form *washed* is one simple word, but the past tense form *has washed* is composed of two simple words. To make these two forms comparable both from a linguistic and a statistical point of view, several options can be considered. One consider *washed* to be improperly spelled: a rational spelling would separate the tense suffix from the root, as in *wash-ed*, so that this sequence would count as two simple words. Alternatively, we might perform a lemmatization of these forms, replacing the forms w*ashed* and *have washed* by a normal form (the infinitive), and attaching to it the grammatical values that differentiate the forms:

---

[3] Which is needed anyway for various reasons: interrogative forms such as *l'a-t-il donnÝ ?* and compound utterances such as: *arc-en-cie*l or *c'est-Õ-dire*.

[4] For example, to solve the ambiguity of *dame* ('give me' vs' ladies').

[5] This body does not deal with technical sublanguages where most compounds are created.

washed = {*wash*, Preterit}, *has washed* = {*wash,* Perfect, 3s}

Lemmatization can be performed automatically on texts by using electronic dictionaries; such an operation modifies the units that are counted.

The notion of formal word excludes meaning, although the search for meaning is usually the main objective of text analysis. However, since words are supposed to carry meaning, dealing with their formal appearance should affect their semantic content in some way. As a matter of fact, if the correspondence between words and meanings were one-to-one, and if it were possible to base the composition of word meanings on the combinatorial (i.e. syntactic) properties of formal words, current processing would provide deep results about the content of texts. However, the empirical situation is not at all that simple:

- first, the correspondence between words and meanings is not one-to-one, since many words are ambiguous;
- second, compositional rules of meaning are not relevant to many strings of words, in other words, there exist numerous compound utterances which have to be recognized since their meaning is independent of their components.

So far, statistical treatment of texts bear on sequences of formal simple words, and ignore the difficulties mentioned. The linguistic tools we will present can automatically correct some of the discrepancies introduced by spelling conventions. We will first discuss the notion of ambiguity, which is a major difficulty arising in automatic text processing.

2. Ambiguity

Words can have a considerable degree of ambiguity, which does not seem to hamper either verbal communication or reading. Only computer parsers suffer from this state of affair, toiling with combinatorial explosions of nonsensical patterns of words, all of which must be rejected. The popular explanation for this paradox is the context effect: when words are put to use in sentences, possibly in relevant extra-linguistic situations, they have only one meaning each. Context is a vague notion which needs to be made precise, so that a computer programme can deal with it. We shall see how the search for the meaningful utterances of a text leads to natural definitions of contexts for simple words. By meaningful utterances, we understand here minimal sequences of words containing ambiguous words and that are defined on a lexico-grammatical basis, that is, on a purely formal basis. As a consequence, they can be recognized by a parser.

The notion of ambiguity is based on the assumption that meaning is located in words. This assumption seems indisputable, at least for words that correspond to concrete entities, such as *bull*, or *bread*.

Let us discuss the meaning of the simple word *bull* and its degree of ambiguity. This word has several literal or proper meanings:

1) the male animal, opposed to the cow;

2) a person with certain features of character. This meaning is said to be a metaphor of meaning 1). A *bull* on the stock market might have a different meaning, hence its own entry;

3) a text promulgated by a pope;

4) the center of a target; etc.
We now consider various uses of *bull* for which it cannot be said that the meaning is carried by the word. Nevertheless, we consider these uses as separate entries:
5) *bull* in the idiomatic sentence: *to take the **bull** by the horns* (which translates word for word in French: 'prendre le taureau par les cornes');
6) the adverbial: *like a **bull** in a china shop*, which translates into French as 'comme un éléphant dans un magasin de porcelaine'. In other words, *bull* here translates as 'éléphant';
7) *a **bull** elephant*;
8) *a **bull** whale*; etc.
We have here 8 words *bull*, each representing an entry in the English lexicon. Hence, the word *bull* is 8 times ambiguous with respect to this lexicon. Talking about the meaning of the first 4 entries makes sense, in particular definitions can be attached to the isolated word. The last 4 entries have no meaning by themselves, for they are part of larger units whose meaning can then be described in the usual way. We call these larger units compound words. Compounds are observed in large numbers for all parts of speech, as we shall see in §3.

But many words have no meaning: grammatical words are generally not considered as carrying meaning. For example, it is difficult to attribute meaning to the prepositions *of* and *to*, although some grammarians have attempted to do so. Even locative prepositions such as *above* and *under* can be shown to have such a wide range of meanings that it is preferable to describe them as empty elements, parts of larger units which have global interpretations that are not specially linked to prepositions. For example, the opposite meanings observed in the sentence pair:

Bob is *for* *any new solution*
Bob is *against* *any new solution*

lead one to attribute meanings to the prepositions, since they are the only words in which the two sentences differ. However, with a different choice of words in the same sentence pattern, the opposition observed above disappears completely:

This drug is *for* *any headache*
This drug is *against* *any headache*

In principle, grammatical words do not belong to the lexicon of the simple words of a language; they are parts of grammar rules, very much as the agent preposition *by* is part of the Passive rule. They can also be part of complex utterances such as compound words and local grammars. We return to their status in §3.

3. Compounds

In order to show how large is the number of compounds that can be found in current texts, we took a sample text from the French journal **Science et Vie** and put in bold characters all the strings we considered to be compounds (cf annex 1).

For *bull*, in the example above, there was one idiomatic or frozen sentence and one frozen adverbial. Our sample text contains others, marked as such:
- **frozen sentences**: qui *portent son nom*, il *met au point*; with support verbs[6] we have: *garde* une *direction fixe* (twice), *donne la direction du nord*, *conserve* son *orientation*, *possède* cette même *propriété*, n'*eut* pas d'*application*;
- **adverbials**: *de près, sans pilote, à grande vitesse, à angle droit, sur sa lancée, plus tard, de haut en bas, de droite à gauche.*

[6] The words that cannot be dissociated from each other are shown in italics. Support verbs are of an auxiliary type subject to limited variations. For example, *avoir, conserver, garder, maintenir* are equivalent in the first four examples (M. Gross 1981).

We observed other compound parts of speech:
- **prepositions**: *jusqu'aux, en passant par*[7]*, par rapport à*;
- **pronouns**: *celle-ci, lui aussi, dans lequel*;
- **determiners**: *des centaines de* N *par* N, *plus de* fidélité *que*, *cette même, tous les*.

The most frequent type of compound is the **compound noun**, as can be seen in the marked text, from the title itself in which we find: *centrale inertielle* and *bouchon de champagne,* to the last *référentiel de guidage*.**Grammatical compounds**

As mentioned above, grammatical words can be seen as semantically empty, but they may be composed of several simple (grammatical) words, and so determine a complex notion of unit. We have the following examples in the text:
- the **clefting** operator: *C'est* en 1850 *que*,
- **negation**: *ne* tourne *pas*, *n*'est *pas*;
- a combination of both: *Ce n'est qu*'avec ... *que*;
- **comparative** terms: *plus de* fidélité *que*, *aussi* sûr *que*.[8]

In order to define units in a coherent way, such dependencies have to be generalized to many other combinations, like the following:
- **governed prepositions**, that is, prepositions that are intrinsic parts of verbs: *aura servi de* guide *à*, *tenir dans*, *réagit à, tend à*, *étend ... au*, *permit au*. Adjectives also govern prepositions: *célèbre pour*, *propres à*; and so do nouns: *direction par rapport à, orientation par rapport à*;
- **governed conjunctions**: *découvre que, constate que, la preuve que*;
- in **Passive** sentences, the auxiliary verb *être* 'be' is correlated with the preposition *par* 'by', as in: *sont* maintenant concurrencés *par*, etc.

The systematic description of all these compounds requires different approaches, depending on the linguistic nature of the compounds.

Frozen **adverbials** need only be listed (M. Gross 1990). Since they are invariable, there is no need for them to be encoded at the morphological level. So far 15,000 compound adverbials have been described in French; this number is to be compared with that of simple adverbs found in current dictionaries: less than 3,000.

Nouns are more complex items than adverbs, since they may vary in gender, number, case. Nouns are stored in a lexicon using inflection codes for the feminine and the plural (B. Courtois 1990). Current dictionaries of French contain about 40,000 simple nouns. So far we have entered about 100,000 compound nouns in the electronic lexicon of French, and our estimate is that it will take 300,000 to 400, 000 compounds to provide a coverage of French similar to that of the 40,000 simple nouns (B. Courtois, M. Silberztein 1990) .

At the other end of the range of complexity, **frozen sentences** require a full lexical and syntactic description, a description not essentially different from that of governed prepositions. Looking at pairs of words such as a verb governing a preposition is limiting to a subcombination of words the needed description of elementary sentences, that is, the full description of the combinations of verbs with their essential arguments. A lexicon-grammar must be built for both free and frozen sentences. It is composed of syntactic tables where the properties of elementary sentences are described (J.-P. Boons, A. Guillet, C. Leclère 1976a. 1976b, A. Guillet, C. Leclère 1992, M. Gross 1968, 1975, 1982; cf. annex 2).

Returning to the examples found in the text, their argument structures, as entered in the

[7] Notice that the following ternary syntactic pattern: *depuis N jusqu'aux N en passant par N* (the *N*s are noun phrases) introduces conjunctional dependencies.

[8] In the last three examples, grammatical compounds are formed of discontinuous (i.e. non connex) strings. This feature requires a more complex parsing procedure than in the case of adverbials and nouns.

lexicon-grammar, are the following[9]:

- frozen sentences (elementary sentences with at least one frozen argument):

N_0 *porter* (*un nom*)$_1$, N_0 *mettre* N_1 (*au point*)$_2$

- free sentences:

N_0 *servir de* N_1 *à* N_2

N_0 *tenir dans* N_1

N_0 *réagir à* N_1

N_0 *tendre à V-inf W, V-inf* is an infinitive form and *W* its complements

N_0 *étendre* N_1 *à* N_2

N_0 *permettre* N_1 *à* N_2, N_0 *permettre à* N_2 (*que S*)$_1$,*S* is 'sentence'

N_0 *constater* (*que S*)$_1$

- adjectival sentences:

N_0 *être célèbre pour* N_1

N_0 *être propre à* N_1.

The lexicon-grammar of free sentences contains about 12,000 such structures, whereas the lexicon-grammar of frozen sentences, still incomplete, contains close to 50,000 structures. Structures are systematically described in terms of the formal operations of substitution (including substitution of the null string, i.e., deletion) and of transformation (including changes in the order of the arguments, as in the Passive). In this way, syntactic tables render explicit the contexts in which a given verb (noun or adjective) may occur, thus providing a general basis for solving ambiguities.

4. Local grammars

There are utterances which, for syntactic and semantic reasons are better described as constituting families of strings rather than lists of independent entries in a lexicon. Families are composed of strings related in meaning. Their members can be entirely synonymous or only share some component of meaning. Moreover, members of a family must share formal properties, that is, subsets of them should have common lexical and syntactic features. We use the term 'local grammar' for the device that represents families, a device formally defined as a finite automaton that can be transformed automatically into a parser of texts (M. Gross, D. Perrin 1989). We will illustrate this general method of representation by means of two examples based on utterances found in our text.

Our first example is the adverbial *comme nous allons le voir plus loin,* 'as we shall see later'. This way of referring to text **ahead** of the adverbial has many formal variants: impersonal pronoun on instead of nous, different locative-temporal adverbs are also possible, as in: *comme nous allons le voir ci-dessous*; such strings are all synonymous and are typical of families described in local grammars. However, we decided to include in the same local grammar adverbials having the same form and function, but referring to text behind the adverbial, hence having a different meaning. The new adverbials are formal variants of the previous ones, since the main changes lie in the tense (future instead of past) and in the choice of locative-temporal adverbs: *comme nous l'avons vu précédemment*, 'as we have seen previously', *comme nous venons de le voir plus haut*, 'as we saw above'.

---

[9] The arguments are subscripted, from '0' for the subject to '3' the maximum number of essential complements. Subscripted *N*s are full noun phrases, free or frozen.

Our second example is that of duration adverbials, of which two occurrences are found in the text:

pendant plus d'un siècle, *pendant plus de 150 ans*

This syntactic form is that of a fairly general noun phrase:

Preposition Determiner Noun

The semantic notion of duration is correlated to specific prepositions[10]: *pendant* and *durant*, 'during', and to nouns that correspond to time units, ranging from fractions of seconds to millenaries; in our examples, we have *siècle* 'century' and *ans* 'years'. The determiners are numerals: *un* 'one', *150*; they are the main source of the differences of meaning between the adverbials of this local grammar. Determiners can be modified both syntactically and semantically by predeterminers like the *plus de* 'more than', seen here. Other similar predeterminers are allowed in the same position: *moins de* 'less than', *presque* 'almost'. Since they are in complementary distribution, that is, grammatically equivalent, they are in the same box. The symbol <E> in the box corresponds to the null string, its appearance means that predeterminers are optional. Other forms of duration adverbials must also be described:

- technical utterances such as *pendant 6 heures 12 minutes et 21 secondes, pendant 6 h. 12 m. 21 s*.,
- more or less idiomatic informal utterances, such as *pendant fort longtemps* (found in the text), *pendant des lustres, pendant des siècles et de siècles,* etc.
- forms that do not use *pendant* or *durant*: some, like *toute la journée* can be analyzed by the deletion of *pendant*, in *pendant toute la journée*, but they will require a separate local grammar, etc.

Hence, a full grammar of duration adverbials has the form of a complex set of graphs each of the type given below (D. Maurel 1990).

These graphs are read from left to right; starting from the initial state (the arrow), all readings must end in the final state (the double square). Each path corresponds to one member of the family, here adverbials of two types.

[10] *Durant* can also be a postposition; pairs such as {*durant 150 ans, 150 ans durant*} can be described by a permutation transformation, but will require a separate local grammar.

5. **Conclusion**

The sample text we have analyzed has undergone a deep transformation of form that brings its content closer to direct perception. Just by underscoring pre-identified compounds and other sequences of words between which strong dependencies hold, we have been able to embed simple words that have no meaning by themselves into larger units that do carry specific meaning. Hence, the counts of simple words, and those of the units marked by our method are quite different:

- from a quantitative point of view. Our sample text is composed of 365 words. If we count the compound words as one unit, the text contains 239 units. Furthermore, if we count the grammatical compounds as units, the text is reduced to 260 units. The reduction of size by one third or more observed on our sample text is a reasonable order of magnitude, confirmed by other similar studies;
- from a qualitative point of view; the objects that are counted and that can be subjected to statistical analysis are very different from the simple words. Further elaboration of the linguistic analysis should accentuate this difference. Meanwhile, the analysis we have presented allows us today to recognize most of the compounds described here.

The analysis presented is indeed operational to a large extent:

- electronic dictionaries of simple and compound words have been built for various languages (Arabic, English, French, German, Italian, Korean, Portuguese and Spanish)[11],
- formalized local grammars of various types have been built for the same languages. Both dictionaries and local grammars can be applied to the parsing of texts in an integrated way by means of the INTEX system developed by M. Silberztein (M. Silberztein 1993, E. Laporte 1994),
- lexicon-grammars, also built for the same languages, have been shown to be directly usable in a parser by E. Roche (1993).

The aim of corpus processing is the study of the distribution of meaning in texts. The analysis we propose here consists of a preliminary series of steps in this direction. The ultimate goal supposes that the problem of identification and localization of meaning has been solved. Linguistics is far from nearing such a stage. However, a realistic linguistic methodology based on Z.S. Harris' theories of language is available and needs only to be applied in order to reach full coverage for most languages. Then, more complete parsers could be developed to include many linguistic features not discussed here (e.g. complex sentences) that would further transform the text, bringing its meaning closer to the surface and at the same time hiding the superficial presentation of content by means of simple words.

---

[11] Lexicon-grammars are being built for Italian (A. Elia 1984; A. Elia, E. d'Agostino, M. Martinelli 1981), for Portuguese (Ranchhod 1990; E. Maceido 1984, J. Malaca-Casteleiro, 1981), and for Spanish (B. Lamiroy 1983, L. Masso-Pellat 1990, C. Subirats 1987). These extensive studies will allow comparisons of Romance languages to be carried out. Many classes of constructions have also been described for English (P. Freckleton 1985; P. Machonis 1988; M. Salkoff 1983), for German (F. Caroli 1984; T. Treigg 1977), for Arabic (M. Chad 1988; M. El Hannach 1988) and for Korean (Hong Chai-Sîng 1984).

REFERENCES

Boons, Jean-Paul; Alain, Guillet; Christian, Leclère 1976a. *La structure des phrases simples en français. I Constructions intransitives*, Geneva: Droz, 377 p.

Boons, Jean-Paul; Alain, Guillet; Christian, Leclère 1976b. *La structure des phrases simples en français, II Constructions transitives*, Paris: Rapport de recherches du LADL, No 6, 85 p., tables et index, 58 p.

Caroli, Folker 1984. Les verbes transitifs à complément de lieu en allemand, *Lingvisticae Investigationes*, Amsterdam-Philadelphia: J. Benjamins B.V. 8.2:225-267.

Chad, Mohammed 1988. Système verbal arabe. Régime des constructions transitives, *Doctoral Thesis*, Université Paris 7: LADL.

Courtois, Blandine 1990. Un système de dictionnaires électroniques pour les mots simples du français, *Langue française* No 87: 11-22, Paris: Larousse.

Courtois, Blandine et Silberztein, Max eds. 1990. Dictionnaires électroniques du français, *Langue française* No 87, Paris: Larousse.

Gross, Maurice 1968. *Grammaire transformationnelle du français. 1-Syntaxe du verbe*, Paris: Cantilène, 188 p.

Gross, Maurice 1975. *Méthodes en syntaxe*, Paris: Hermann, 412 p.

Gross, Maurice 1981. Les bases empiriques de la notion de prédicat sémantique, Formes syntaxiques et prédicats sémantiques, A. Guillet et C. Leclère eds., *Langages*, No 63: 7-52, Paris: Larousse.

Gross, Maurice 1982. Une classification des phrases figées du français, *Revue québécoise de linguistique*, Vol. 11, No 2: 151-185, Montreal: Presses de l'Université du Québec à Montréal.

Gross, Maurice 1990. *Grammaire transformationnelle du français. 3-Syntaxe de l'adverbe*, Paris: ASSTRIL, 670 p.

Gross, Maurice, Dominique Perrin eds. 1989. *Electronic Dictionaries and Automata in Computational Linguistics*, Lecture Notes in Computer Science, Berlin: Springer Verlag, 110 p.

Guillet Alain, Leclère Christian, 1992. *La structure des phrases simples en français: constructions transitives locatives*. Droz: Geneva, 445 p.

El Hannach, Mohammed 1988. Syntaxe des verbes psychologiques de l'arabe, *Doctoral Thesis*, Université Paris-7: LADL.

Elia, Annibale 1984. *Le verbe italien, Les complétives dans les phrases à un complément*, Bari-Paris: Schena-Nizet, 305 p.

Elia, Annibale, Emilio D'Agostino, Maurizio Martinelli 1981. *Lessico e strutture sintattiche. Introduzione alla sintassi del verbo italiano*, Naples: Liguori.

Freckleton, Peter 1985. Une comparaison des expressions de l'anglais et du français. *Doctoral Thesis*, Université Paris 7: LADL.

Harris, Zellig S. 1952. Discourse Analysis, *Language* 28, Baltimore: The Waverly Press, pp.1-30.

Harris, Zellig 1988. *Language and Information*, New York: Columbia University Press, 119 p.

Hong, Chaï-sông 1984. *La classe des verbes de mouvement en coréen contemporain*, Lingvisticae Investigationes Supplementa, Amsterdam Philadelphia: J. Benjamins B.V., 309 p.

Laporte, Eric 1994. Experiments in Lexical Disambiguation Using Local Grammars, *Papers in Computational Lexicography* (COMPLEX), Budapest: Research Institute for Linguistics,

Hungarian Academy of Sciences, pp.163-172.
Lamiroy, Béatrice 1983. *Les verbes de mouvement en français et en espagnol. Etude de syntaxe comparée de leurs infinitives*, Lingvisticae Investigationes Supplementa, Amsterdam-Philadelphia: J. Benjamins B.V., 309 p
Maceido-Oliveira, Elisa 1984. *Syntaxe des verbes psychologiques du portugais*, Instituto Nacional de Investigaçào C□entifica, Centro de L□nguistica da Universidade de Lisboa: Lisbon, 198 p.
Machonis, A. Peter 1988. Support Verbs: An Analysis of **be Prep X** idioms, *The SECOL Review*, 12.2: 95-125.
Malaca Casteleiro, Joao 1981. *Sintaxe transformacional do adjectivo*, Instituto Nacional de Investigaçào C□entifica, Centro de L□nguistica da Universidade de Lisboa: Lisbon, 561 p.
Maurel, Denis 1990. Adverbes de date: étude préliminaire à leur traitement automatique, *Lingvisticae Investigationes* XIV:1, Amsterdam-Philadelphia: J.€Benjamins Pub. Co., pp. 31-63.
Pellat-Masso, Luisa 1990. Une description formelle des expressions figées de l'espagnol, *Mémoires du CERIL* 5:22-290.
Ranchhod, Elisabete 1983. On the Support Verbs **ser** and **estar** in Portuguese, *Lingvisticae Investigationes*, Vol. VII, No 2: 317-353, Amsterdam-Philadelphia: J. Benjamins B.V.
Ranchhod, Elisabete 1990. *Sintaxe dos Predicados Nominais com **Estar*,** Instituto Nacional de Investigaçào C□entifica, Centro de L□nguistica da Universidade de Lisboa: Lisbon, 477 p.
Roche, Emmanuel 1993. Une représentation par automate fini des textes et des propriétés transformationnelles des verbes, *Lingvisticae Investigationes* XVII:1, Amsterdam Philadelphia: J.€Benjamins Pub. Co., pp. 189-222.
Salkoff, Morris 1983. Bees are swarming in the garden, *Language*, Vol. 59, No 2, Baltinore: The Waverly Press, pp. 288-346.
Silberztein, Max 1993. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Paris: Masson, 233 p.
Subirats-Ruddeberg, Carlos 1987. *Sentential Complementation in Spanish, A lexico-grammatical study of three classes of verbs*, Lingvisticae Investigationes Supplementa, No 14, Amsterdam-Philadelphia: J. Benjamins Pub. Co., 290 p.
Treig, Thomas 1977. Complétives en allemand. Classification, *Rapport de recherches du LADL*, 7:39-203.

ANNEX 1

UNE CENTRALE INERTIELLE DANS UN BOUCHON DE CHAMPAGNE

par RENAUD DE LA TAILLE : SCIENCE ET VIE No 896 MAI 92

Pendant plus d'un siècle, une lourde toupie lancée à **des centaines de tours par seconde** aura servi de guide à **tous les** mobiles avec ou **sans pilote**, depuis la torpille de **la première guerre jusqu'aux** sondes qui sont allées filmer Saturne **de près**, **en passant par** les avions. Mais ces gros gyroscopes sont maintenant concurrencés par un minuscule système **à diapason** qui tient dans **un bouchon de champagne**.

C'est en 1850 que le Français Léon Foucault, physicien célèbre pour son pendule et pour les courants qui **portent son nom**, invente le gyroscope. En étudiant l'étonnante stabilité d'une toupie lancée **à grande vitesse**, il découvre que **celle-ci** réagit **à angle droit** à toute sollicitation et tend à **garder une direction fixe par rapport aux étoiles**. **Sur sa lancée** il **met au point** le gyrocompas qui, dans un **navire en fer**, **donne la direction du nord** avec **plus de** fidélité **qu'**une boussole.

Il étend **plus tard le champ de ses recherches** au **plan d'oscillation** d'un pendule, qui, **lui aussi**, conserve son orientation **par rapport aux étoiles** fixes, puis aux **systèmes oscillants** quelconques qui possèdent **cette même** propriété -- si on fait rouler sur **un bord de table** un mince **fil d'acier** dont l'extrémité vibre **de haut en bas** ou **de droite à gauche**, on constate que **le plan de vibration ne** tourne **pas** lorsqu'on fait rouler la tige, mais qu'il reste dans le plan **dans lequel** sa vibration a été lancée, qu'il soit vertical, horizontal ou oblique.

L'expérience du **pendule de Foucault**, qui permit aux spectateurs de voir littéralement la Terre tourner, **apportait la preuve** qu'**un mouvement oscillant gardait une direction fixe par rapport à un système de référence** constitué par les étoiles. Mais cette propriété **n'**eut **pas d'application pratique pendant plus de 150 ans**. **Ce n'est qu'**avec l'avènement des microstructures propres à **l'industrie des composants électroniques**, que le **système oscillant** constituera, **comme nous allons le voir plus loin**, le moyen **le plus** pratique d'avoir **un référentiel de guidage aussi sûr que** le gyroscope, mais considérablement miniaturisé.

ANNEX 2